Pig Introduction

|  |  |
| --- | --- |
| **Objective:** | This demo executes a simple iterative process -  K-Means Clustering - using embedded Pig. |
| **Location of Files:** | ~/materials/data |
| **Successful Outcome:** | The Pig Latin in turn invokes a Java-based UDF to compute the nearest centroid for each data point. |
| **Eclipse Project** | pig\_intro |
| **Exercise directory** | ~/workspace/pig\_intro |

1. Put the test data set student.txt into your home directory in HDFS

2. Run one of the following commands to execute the process:

Pig with MapReduce:

PIG\_OPTS=“-Dpython.cachedir.skip=true" pig kmeans.py

Local mode will by far be the fastest option with the small sample data set, but the interesting comparison is between Pig on MR and Pig on Tez. Expect to see Pig on Tez perform 3-5x faster than Pig on MR. The larger and more iterative the process, the more significant this performance boost will be, due to session/container re-use and in-memory caching performed by Tez.

**END**